Exponential Language Models, Logistic Regression, and Semantic Coherence

Can Cai, Ronald Rosenfeld, Larry Wasserman [*]
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
ccai@stat.cmu.edu,roni@cs.cmu.edu,larry@stat.cmu.edu

June 20, 2000

Abstract

In this paper, we modify the traditional trigram model by using utterance-level semantic coherence features in an exponential model. The semantic coherence features are collected by measuring the correlations among content-word pairs occurring in sentences of two corpora, the real corpus and a corpus generated by the baseline trigram model. The measure we use for estimating the semantic association of content word pairs is Yule's Q statistic. For our preliminary analysis, we have further simplified the modeling task by extracting a small set of statistics from each sentence-based Q statistics and applying them as features to the exponential model. We also simplified the process of obtaining the MLE solutions of the exponential models by approximating it with a logistic regression model. We account for the uncertainty in the estimates of Q by constructing confidence intervals. The new model results in a slight reduction in test-set perplexity. We also discuss and compare alternative measures of associaztion, such as χ statistics.

1 Introduction

Trigram models decompose the probability of a sentence into a product of conditional probabilities by the chain rule:

$$\begin{aligned} P(s) &= \prod_{i=1}^{k} P(w_i|w_1\cdots w_{i-1}) \\ &\approx \prod_{i=1}^{k} P(w_i|w_{i-n+1}\cdots w_{i-1}) \qquad (1) \end{aligned}$$

One disadvantage of this approach is that global sentential information is hard to express. In particular, linguistic features cannot be easily integrated into the model.

Rosenfeld (1997) introduced a Maximum Entropy model which directly models the probability of an entire sentence. According to this model, the probability of a given sentence s is

$$P(s) = \frac{1}{Z}\cdot P_0(s)\cdot exp(\sum_i \lambda_i f_i(s)) \qquad (2)$$

where $P_0(s)$ is a baseline probability of a sentence s, Z is a normalizing constant, $f_i(s)$ are *features*, or arbitrary computable properties of the sentence (such as the length of the sentence, the number of verbs in the sentence, etc.), and λ_i are the associated weights. The maximum likelihood estimates of $\hat{\lambda}_i$ of λ_i satisfies

$$E_{\hat{\lambda}}(f_i(s)) = \frac{1}{N}\sum_s f_i(s) \qquad (3)$$

where N stands for total number of sentences in the corpus. The essential point here is to find powerful features that significantly modify the baseline estimate, to make it better conform to actual language. In previous work, Zhu et al (1999) used the whole sentence maximum entropy model with linguistic features that captured variable-length syntactic sequences sentence, resulting in a slight reduction in perplexity on the Switchboard corpus.

In this paper, we explore a complemantary aspect of language, namely whitin-sentence semantic coherence. The traditional trigram model does not capture semantic correlations among content words, particularly when they are far apart. Once these deficiencies of the trigram model are quantified, they can serve to define new features.

In Section 2, we describe how we measured the semantic association of content word pairs within a sentence by using Yule's Q statistic. Section 3 introduces a new

[*] We are very grateful to Stanley F. Chen and Xiaojin Zhu for much appreciated help and discussions.

and more convenient procedure for fitting the exponential model with new features. Sections 4 presents preliminary results from the new models. Section 5, which is an extension of section 2, discusses the confidence interval of Q statistics. Section 6 briefly discusses the alternative use of χ statistics, and lists future work. All experiments were run on the Broadcast News corpus.

2 Semantic Associations

One way to design semantic coherence features is to look for correlations in the target corpus which are not predicted by the baseline trigram. To this end, we generated a corpus of 'pseudo-sentences' from a Modified Kneser-Ney trigram model (Chen and Goodman 1998) trained on 103 million words of Broadcast News data. We then extracted content word pair counts from the original training corpus, as well as from the corpus of 'pseudo sentences'. Since the trigram model captures local correlations well, we decided to focus on the remote correlation between two content words. Therefore, the content word pairs studied in our analysis are only those where the two content words co-occurring in the same sentence with at least five words in between them. For pragmatic reasons we defined 'content words' to be the most common 50,000 words in the training data, but excluding the most common 200 words.

2.1 Contingency Table

Next, we built a contingency table for each content word pair. The measure of semantic association will be defined based on the contingency table. For a content word pair (Word1 Word2), the contingency table is

		Word1	
		Yes	No
Word2	Yes	C_{11}	C_{12}
	No	C_{21}	C_{22}

where C_{11} is the count of sentences in the training corpus which contain both words (with at least 5 words between them); C_{12} is obtained by subtracting C_{11} from the number of sentences with Word1 in it; C_{21} is the obtained by subtracting C_{11} from the number of the sentences with Word2; C_{22} is the total number of sentences minus the other three counts. Since we only consider content word pairs separated by at least 5 words (to exclude "trigram effects"), C_{12} and C_{21} overlap at the set of sentences in which both Word1 and Word2 occur but with fewer than 5 word apart. Because of this, the table as defined above is not a contingency table in the usual sense.sense.

2.2 Yule's measure of association

After building the contingency table for each of the content word pairs, a proper measure of association must be chosen. We decided to use Yule's measure of association, which is also called Yule's Q statistic,

$$Q = \frac{C_{11}C_{22} - C_{12}C_{21}}{C_{11}C_{22} + C_{12}C_{21}}. \tag{4}$$

The higher the value of Q, the stronger the correlation between the two words.

There are several reasons for choosing Yule's Q. First, unlike many other measures which can only take positive values, the value of Q range from -1 to 1. This is important because the negative values can tell us to which extent the two content words tend not to occur together. The second advantage of Q is that it is well defined even for contingency tables with the cell counts equal to zero, which isn't true for some otherwise reasonable measures, like mutual information:

$$\hat{I} = \sum_i \frac{C_{ij}}{N} \log N \frac{C_{ij}}{C_{+i}C_{+j}}, \tag{5}$$

The Mutual Information statistic is undefined if there exists a cell with count equal to zero in the contingency table. Third, for our data, the Q statistics tend to assume the full dynamic range $[-1, 1]$. Some other measures, for example, the correlation coefficient statistics ρ, which has the form

$$\hat{\rho} = \frac{C_{11}C_{22} - C_{12}C_{21}}{\sqrt{C_{1+}C_{2+}C_{+1}C_{+2}}}, \tag{6}$$

has a very narrow dynamic range in practice.

The following are some examples of Q values of some content word pairs calculated based on the 103 million word training corpus:

Table1. Q values of some content word pairs

ENTERTAINMENT WITNESS	-0.89
ANGELES CONGRESS	-0.70
ABORTION REPUBLICAN	0.74
ARTISTS SONY	0.89
M H	1

These Q's are indeed consistent with our intuition about the semantic correlation of these content word pairs[1].

2.3 Q statistic for sentences

For any given sentence, we can now calculate a list of Q statistics for all the content word pairs in it, based

[1] One may be surprised to see $Q(M, H) = 1$. This is because in our training data, the only sentences with either M or H as words contain the sequence "M * A * S * H “.

on the contingency table with counts collected from the training corpus. Since we are interested in the semantic coherence of the sentence as a whole, we further calculate a set of descriptive sentence-level statistics from that list. The statistics we use are the mean, median, maximum and minimum values in the list. The following diagram illustrate our process for deriving sentence-level statistics:

$$\begin{tikzcd}
\text{A sentence} \arrow[d, Rightarrow] \\
W_1W_2,\, W_1W_3, \cdots \arrow[d, Rightarrow] \\
Q_{12}, Q_{13}, \cdots \arrow[d, Rightarrow] \\
Q_{\text{mean}}, Q_{\text{median}}, Q_{\text{max}}, Q_{\text{min}}
\end{tikzcd}$$

W_iW_j represent content word pairs and Q_{ij} is their associated Q value. Repeated word pairs are counted only once. Figures 1 and 2 show, in order, the histograms of the mean, medium, maximum and minimum of Q statistics for a 59929-sentence corpus[2] of Broadcasting News data, and an equal number of sentences generated from the baseline model.

Mean Median

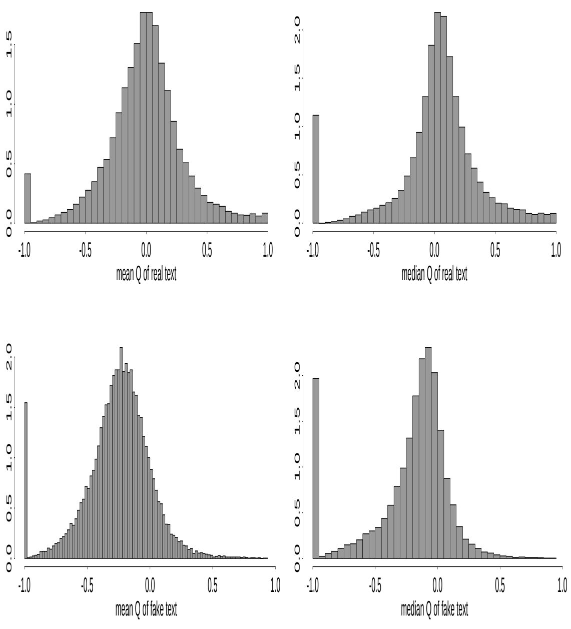

Figure1

Maximum Minimum

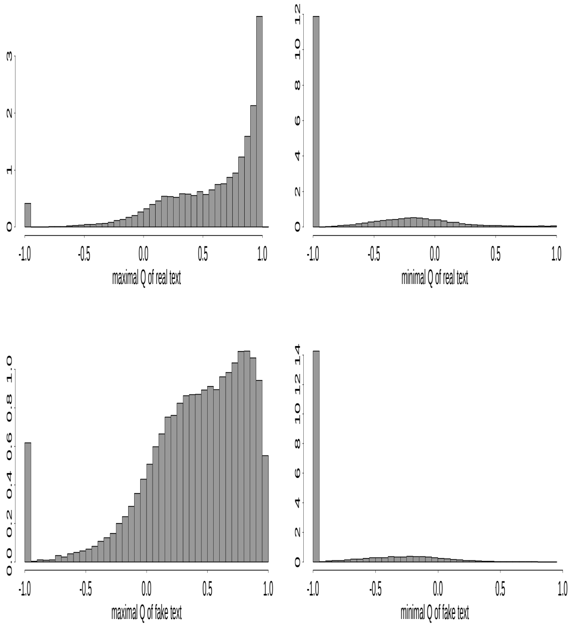

Figure2

For each of the statistics, the top graph is for the Broadcast News sentences [3] and the bottom one is for sentences generated from the baseline model. Comparing the top and bottom histograms of each statistic, we note that there are obvious differences between the distributions of the Q statistic of the real sentences and pseudo sentences. Especially for the histograms of Q_{mean} and Q_{median} the centers of the distributions of Q_{mean} and Q_{median} of the pseudo sentences are to the left of those of the original sentences. Thus pairwise correlation are weaker in the pseudo corpus. This is of course consistent with our intuition that the baseline trigram model does not adequately capture within-sentence semantic correlations.

3 Fitting the Whole Sentence Entropy model

Because of the clear distributional differences of these four statistics between the two corpora, we decide to apply them as features to the whole sentence exponential model of equation 2.

The critical procedure of fitting the model is to find the maximum likelihood solution of λ_i. This is a nontrivial problem in exponential models in general, and is particularly problematic in whole sentence models. Chen and Rosenfeld (1999) provides an approach to obtaining the MLE of λ_i by using the Generalized Iter-

[2] This corpus is not among the training data used to train the baseline trigram model

[3] In the histogram, we excluded the sentences without any content word pairs, for which all the four Q statistics are zero.

ative Scaling algorithm with the support of statistical sampling methods. While this procedure was shown to be computational feasible for practical construction of competitive language models, it is still too computationally intensive to be used for model selection. In this section we describe how to fit the whole sentence model by formulating a related logistic regression problem.

3.1 Logistic Regression Models

There are three major advantages to using logistic regression for our problem. First, it is much easier to fit than the exponential model; there are full-featured statistical software packages for regression. Second, logistic regression provide us a more efficient and convenient way to do model selection when there are hundreds or thousands of features. The third advantage is that the logistic reformulation enables us to do non-parametric regression.

To formulate the whole sentence model fitting as a logistic regression problem, we define a new variable Y such that

$$Y = \begin{cases} 1 & \text{if} \quad s \in P \\ 0 & \text{if} \quad s \in P_0 \end{cases}$$

where $s \in P$ means the sentence came from the original corpus, and $s \in P_0$ means the sentence came from the pseudo corpus.

We also define a new function $h(s) = P(Y = 1|s)$. By Bayes Theorem:

$$\begin{aligned} h(s) &= P(Y=1|s) \\ &= \frac{P(s|Y=1)P(Y=1)}{P(s|Y=1)P(Y=1) + P(s|Y=0)P(Y=0)} \end{aligned} \tag{7}$$

By design, we choose $P(Y = 1) = P(Y = 0) = \frac{1}{2}$. Thus,

$$h(s) = \frac{P(s)}{P(s) + P_0(s)}.$$

Hence,

$$\frac{h(s)}{1 - h(s)} = \frac{P(s)}{P_0(s)}. \tag{8}$$

Substituting $P(s)$ with the right hand side of equation (2), we get

$$\begin{aligned} \frac{h(s)}{1-h(s)} &= \frac{Z^{-1} P_0(s) \exp(\sum_i \lambda_i f_i(s))}{P_0(s)} \\ &= \frac{1}{Z} exp(\sum_i \lambda_i f_i(s)). \end{aligned} \tag{9}$$

By taking logarithm of both side of (3), we obtain

$$\begin{aligned} \log\left(\frac{h(s)}{1-h(s)}\right) &= -\log Z + \sum_i \lambda_i f_i(s) \\ &= \beta_0 + \sum_i \beta_i f_i(s) \end{aligned}$$

where $\beta_0 = -\log Z$ and $\beta_i = \lambda_i$.

If we let $f_i(s) = x_i$ in formula (10), it gives us the exact form of logistic regression (Generalized Linear Models, or GLM)

$$\text{logit}(s) = \beta_0 + \sum_i \beta_i x_i. \tag{10}$$

The probability of a sentence as estimated by this model will be

$$\hat{P}(s) = P_0(s) exp(\hat{\beta}_0 + \sum_i \hat{\beta}_i x_i) \tag{11}$$

3.2 Generalized additive models

The features we apply to the model do not necessarily have linear relationships with logit(s). Actually, in most of the cases, they have non-linear relationships. Therefore, instead of estimating the coefficient β_i which corresponds to the λ_i values in the maximum entropy model, we can estimate a smooth function of x_i by fitting a Generalized Additive Model (GAM)

$$\text{logit}(s) = s_0 + \sum_i s(x_i), \tag{12}$$

which is also called nonparametric logistic regression. In this work, we use the smoothing spline as the smooth function for each feature. In the nonparametric case, the probability of the sentence is estimated as:

$$\hat{P}(s) = P_0 exp(\hat{s}_0 + \sum_i \hat{s}(x_i)) \tag{13}$$

4 Preliminary results

To train our GLM and GAM models, we used the two corpora mentioned above: a sample of 59,926 real sentences from the Broadcasting News domain, and a same-size corpus of pseudo-sentences, generated from the baseline trigram model $P_0()$. We fit the logistic regression models using six features: the mean, median, maximum and minimum of the Q values of the content word pairs in a given sentence, plus the length (number of words) and number of content word pairs in each sentence. The reason for the latter two features is the slight but systematic difference we observed in the average sentence length (and consequently in the number of content word pairs per sentence) between the original corpus and the pseudo corpus.

After fitting the models using standard statistical software packages, we observed that all six features were statistically significant in terms of χ^2 tests. We then set out to measure the effect of the features on perplexity. Unlike conventional conditional language models, in whole sentence models perplexity cannot be computed analytically, because the normalizing constant Z cannot be so computed. However, it can be estimated to any arbitrary accuracy. Using the technique described in (Zhu et al, 1999), we measured sentence-level and word-level perplexity of both the GLM and the GAM models, using a yet-unseen test set of 56697 sentences, and compared it to the baseline. Results are sumamrized in table 2.

Table2. Perplexity of GLM, GAM and the baseline model

model	perplexity	% reduced
baseline	111.43	
GLM model	109.78	1.5%
GAM model	107.55	3.5%

Although the reduction in perplexity at the word level is slight, it has been achieved with only 6 new parameters. Since the parameters are applied at the sentence level, their average effect at the word level is much smaller (at a sentence level, they result in an improvement of 21% and 43%, respectively, in the average sentence log-likelohood).

5 Confidence intervals of Q statistics

Like all other point estimations, the estimator of Q statistics expressed in (4) does not take into account sample variations. Take the cases when $C_{11} = 0$ as an example, if a content word pair never occurs in the training corpus, the estimate of its Q value is -1, regardless of the marginal counts of the two words. It may not be an accurate estimate, especially when the marginal counts of the words are small; this content word pair may well occur in new data.

The statistic Q is an estimate of a population parameter θ. To account for the uncertainty in Q we compute a confidence interval. The greatest statistical variation is in C_{11} so we begin by computing a confidence interval for $p = E(C_{11})/n$, treating C_{12}, C_{21} and C_{22} as constants. Specifically, note that

$$C_{11} \sim \text{binormial(N, p)}, \qquad (14)$$

where N is the total number of sentences in the training corpus and p is the probability of two words occurred together in the same sentences and at least five words in between them. Then an exact $1-\alpha$ confidence interval for p is

$$\frac{1}{1+\frac{N-C_{11}+1}{C_{11}}F^{-1}_{2(N-C_{11}+1),2x,\alpha/2}} \le p \le \frac{\frac{C_{11}+1}{N-C_{11}}F^{-1}_{2(C_{11}+1),2(N-C_{11}),\alpha/2}}{1+\frac{C_{11}+1}{N-C_{11}}F^{-1}_{2(C_{11}+1),2(N-C_{11}),\alpha/2}} \qquad (15)$$

where $F^{-1}v_1, v_2, \alpha$ is the upper α cutoff from an F distribution with v_1 and v_2 degrees of freedom (Clopper and Pearson, 1934). The confidence interval of C_{11} is just $[N\hat{p}_{low}, N\hat{p}_{upper}]$, where $\hat{p}_{low}$ and $\hat{p}_{upper}$ represents the lower and upper bound of the $1-\alpha$ confidence interval of p. Substituting the upper and lower bound of C_{11} into (4), we can get the $1-\alpha$ confidence interval for Q. We use the four descriptive statistics of the upper and lower bound of Q values as new features together with the statistics of the point estimates of Q.

Also, by observing the confidence intervals of Q for content word pairs from both the real and pseudo corpora, we notice two things. First, the lengths of the confidence intervals tend to be shorter for the content word pairs in real corpus than those in the pseudo sentence corpus. Figure3 shows the distributions of the utterance mean of the lengths of the confidence interval for the Q values in a real and a pseudo corpus with 59929 sentences each. The histogram on the top is for sentences from the real corpus while the bottom one are for those from the pseudo corpus. The vertical solid line indicates the medians of the distributions. As can be seen from the graph, the distribution of the utterance mean length of the confidence interval of Q from the real corpus is more right-skewed than that from the pseudo corpus.

Second, in the real corpus, most of the short confidence intervals of Q, for example, the intervals with length less than 0.2, are shifted towards 1. But for the pseudo corpus, those short intervals are mostly centered on 0; also, the number of those short intervals is much smaller in the pseudo corpus than in the real corpus.

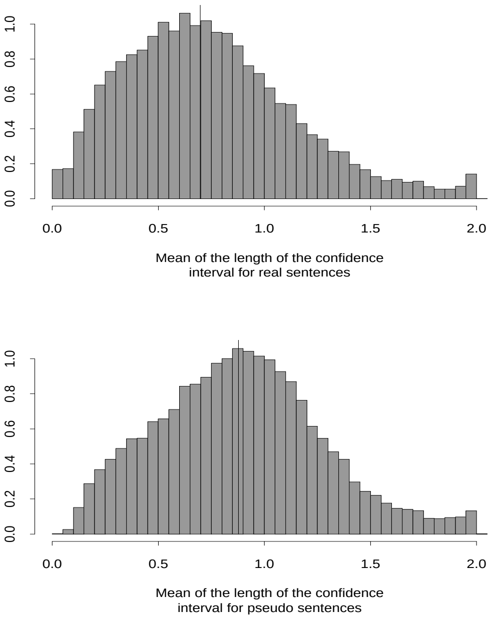

Figure3
Utterance mean of the lengths of the confidence interval of Q

Based on these findings, we add the following new features to our GLM and GAM models: the number of content word pairs whose Q values has a confidence interval shorter than 0.2, the four descriptive statistics of the length of the confidence interval of Q in addition to the same four statistics for the point estimate of Q, upper and lower bound of Q. After adding these new features, word-level perplexity improvement is slightly increased, to 3% for the GLM model, and 5% for the GAM model.

6 Further discussion

6.1 χ statistics

Besides using the Q statistic as a measure of association, we also considered another measure, χ statistics, which is defined as

$$\chi_{ij} = \frac{C_{ij} - E_{ij}}{\sqrt{E_{ij}}} \tag{16}$$

(DuMouchel, 1999) Where C_{ij} is the actual counts of a content word pair $(\mathrm{word_i}, \mathrm{word_j})$, and E_{ij} is their expect counts of occurrence under the independence assumption. The advantage of Chi over Q is that it is scaled by the standard distribution of C_{ij}, because one can approximate the distribution of C_{ij} as a $\mathrm{poisson(E_{ij})}$ with variance equal to E_{ij}. Our preliminary study of χ shows that the GLM and GAM with the four descriptive statistics of χ together with sentence length and number of content word pairs as features reduces the perplexity by 4–5% for both models. The improvement that this new statistics brings merits further study.

6.2 Other issues

The results reported here are very preliminary. Our goal is to built a small set of features that will capture global semantic coeherence. Shannon-style experiments show that if such a feature is even half as good as human judgment, perplexity reduction will be substantial. The main difficulty we are still facing is how to model directly the distribution of the entire set of content words, as opposed to modeling of individual word pairs, which is what was attempted here.

Even within the pairwise-correrlation approach, we are still looking for more and better features to add to our model. Also, we would like to assess how efficient logistic regression models are in approximating the maximum likelihood solution to the whole sentence exponential model. We can learn that by calculating the relative efficiency of the estimators of logistic regression to the MLE of λ_i. In addition, the solution provided by the regression models can be used as a starting point for calculating the MLE.

References

[1] Yvonne M. M. Bishop and Stephen E. Fienberg and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass., The MIT Press,1975

[2] George Casella and Roger L. Berger. *Statistical Inference* Belmont, California, Duxbury Press, 1990.

[3] Stanley F. Chen and Joshua T. Goodman. *An Empirical Study of Smoothing Techniques for Language Modeling.* Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[4] Stanley Chen and Ronald Rosenfeld. *Efficient Sampling and Feature Selection in Whole Sentence Maximum Entropy Language Models.* Proc. ICASSP'99, Phoenix, Arizona, March 1999.

[5] C. J. Clopper and E. S. Pearson. *The use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.* Biometrika 26, 404-413, 1934.

[6] William DuMouchel. *Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System.* The American Statistician, Vol. 53, No. 3, page 177-202, August 1999.

[7] Stephen Fienberg. *The Analysis of Cross-Classified Categorical Data, Second Edition.* Cambridge, Mass., The MIT press, 1990

[8] Ronald Rosenfeld.*A Whole Sentence Maximum Entropy Language Model, Proc. IEEE workshop on Speech Recognition and Understanding, Santa Barbara, California, December 1997.*

[9] Xiaojin Zhu, Stanley Chen and Ronald Rosenfeld. *Linguistic Features for Whole Sentence Maximum Entropy Language Models.* Proc. Eurospeech'99, Hungary, September 1999.